Lecture 9 presents details of structure-structure alignments and various web-based servers for their calculation. Alignments are important for structural annotation of new crystal structures. A typical query is - Is the newly determined structure unique or simply a structure of a well-recognized protein fold? Moreover, these methods are important for protein structure predictions.

Let us run through an exercise to anchor our understanding of structural alignments and the mechanics of their computation.

Topic 1

Go to rcsb.org and search the PDB for a crystallographic structure that is not currently annotated in SCOP, CATH or Dali (or other fold library). Perhaps check the newly released entries in PDB for a structure or implement an alternative search of PDB. Use structure alignment methods to propose a fold type for your selected structure. For calculation of the alignments, you will most likely have to download the structure to your computer (namely, Desktop) and then upload the structure to a web-based server for a comparison with known structures (a less time consuming option is to use a PDB subset of non-redundant topological folds).

Use Dali, FATCAT and or other methods (see, e.g., http://en.wikipedia.org/wiki/Structural_alignment_software) that allow the uploading of a structure to their server. Do not use sequence alignment techniques to search for the fold classification. The focus is on protein folds and the search of structural databases for 3D comparisons.

Because there are many available structures to explore, try not to duplicate the work of a posted result by a classmate. Keep in mind, however, alignment calculations may take some compute time, so use your time judiciously.

Report the following:

Describe your selected protein in terms of its functional annotation (if known), resolution and the number of residues.A suggested fold type and the RMSD values calculated from the alignments of your protein with other proteins and their sequence identities – namely, the output from Dali, FATCAT, etc. Note that is there are many different computational schemes for calculating structure-structure alignments of proteins and their scoring methods (see, e.g., http://en.wikipedia.org/wiki/Structural_alignment).A molecular-graphics image of an alignment between your selected protein and the calculated homologue or structure neighbor (the so-called remote homologue).

Crystallographic structure: 6N7Z
6N7Z has a residue count of 375 and only a single unique protein chain. It is a thienopyrimidine-based allosteric inhibitor of human farnesyl pyrophosphate synthase and is classified as a transferase/transferase inhibitor.
Structure Alignment methods:

Download 6N7Z pdb file to upload to DALI for structural alignment against any other known protein structures. The top two results of structural neighbors are 6ely-B and 2vlc-A. As seen later on in FATCAT, 6ely-B is not structurally similar when examined in a pairwise alignment while 2vlc-A is. Hence forth, I shall only be comparing 6N7Z to 2vlc-A. According to the DALI alignment, 2vlc-A has a rmsd of 1.3 with a 67% sequence identity.

No:	Chain	Z	rmsd	lali	nres	%id	PDB	Description
1:	6ely-B	39.2	1.3	260	264	61	PDB	MOLECULE: MISTLETOE LECTIN I;
2:	2vlc-A	36.1	1.3	262	518	67	PDB	MOLECULE: TYPE 2 RIBOSOME-INACTIVATING PROTEIN CINNAMOMIN

Above is the Dali Alignment of the portion of 6N7Z that aligns with 2vlc-A. According to this alignment, the 6N7Z has relatively long regions of extended sheets interspersed with some coils and very few helices located right before position 500 of the alignment.

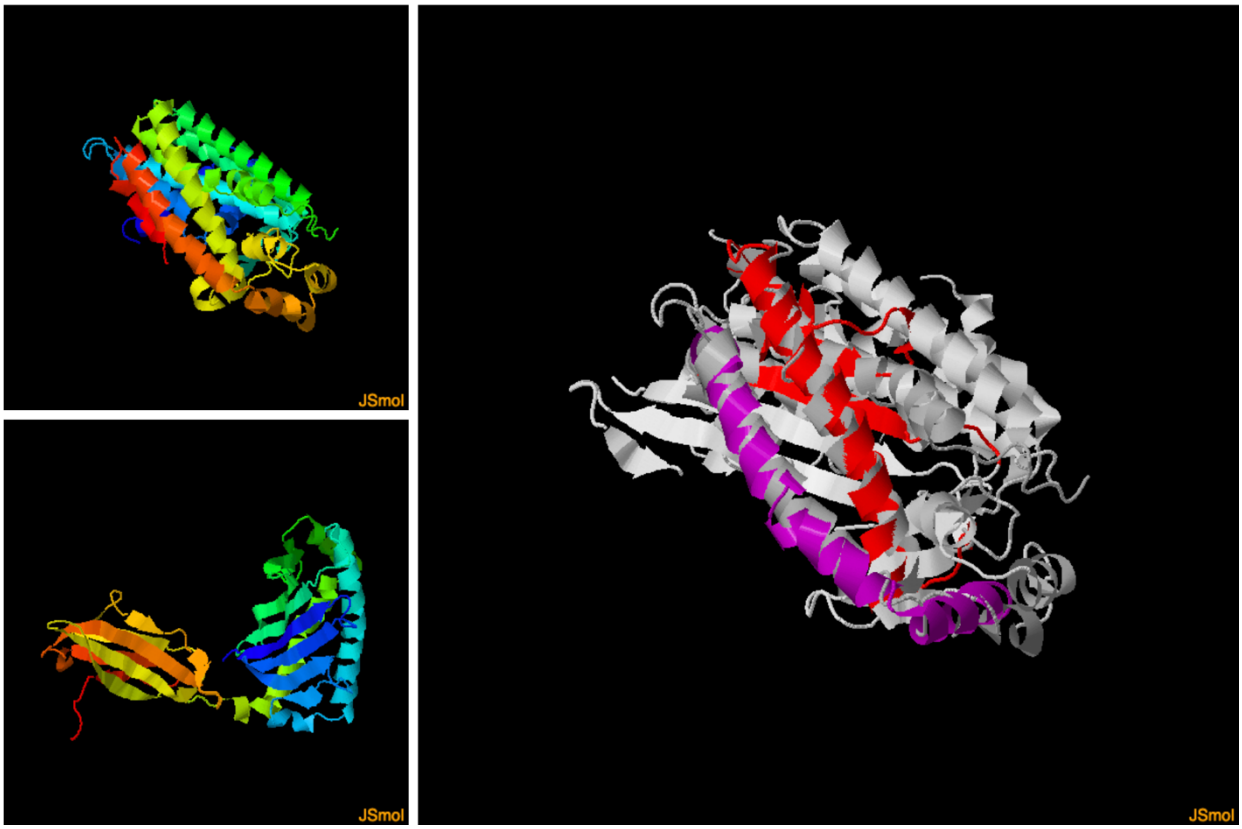
Go to RCSB and view structure similarities. No structure alignment is available for 6N7Z.F, but is represented by 4NUA.A which has > 95% sequence identity and is the crystallographic structure for the effects of Lysine 200 and Phenylalanine 239 on FPPS. Interestingly, although 4NUA.A has a >95% sequence identity, it did not appear in the structural alignments at all from DALI.

Use FATCAT to do a pairwise comparison between 6N7Z pdb file and 6ely-B and 2vlc-A. According to FATCAT pairwise structural alignment, 6ely-B is NOT significantly similar to 6N7Z with a p-value of 9.98e-01.

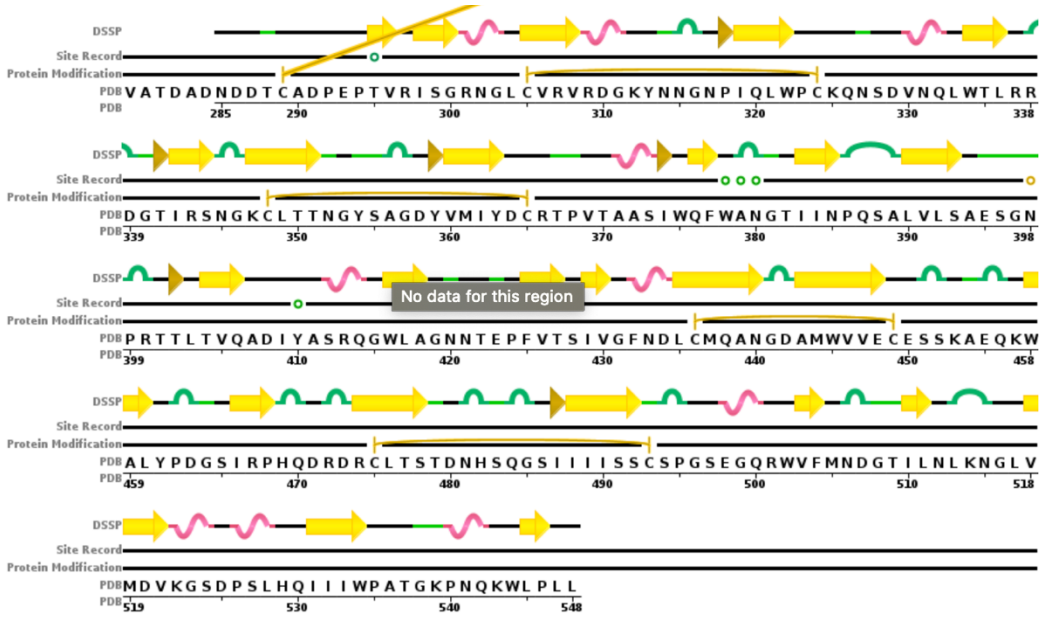
However, 2vcla-A IS significantly structurally similar to 6N7Z with a p-value of 4.75e-02. There are 97 equivalent positions with RMSD of 3.18 and 1 twist

JSmol display: str1 (left top); 2clva (left bottom); superposition (right), str1 in gray, 2clva in different colors (seperated by twists)

Synchronize structure windows



Above is a superimposed image (right) of 6N7Z (top left) over the structurally similar 2vlc-A (bottom left).



Pictured above is the a sequence chain representation of the 2vlc-A sequence starting at around 280, where the overlap between 6N7Z occurs with 2vlc-A. Present in 2vlc-A are numerous beta strands and 3/10 helices. Since our highest structural alignment sequence ID was 2vlc-A from DALI and the pairwise structural alignment returned with high structural similarity from FATCAT, a good prediction of the fold type for 6N7Z can come from 2vlc-A. When searching SCOP for 2vlc-A, there were no proteins matching it, however, the entry of 2vcl-A is represented by 2VGR-C. Following 2VGR-C, I found that they are classified as a protein fold of: Ferredoxin-dependent bilin reductase-like. Extrapolating from this data, my best hypothesis is that 6N7Z is similar to 2vcl-A and can be considered with having a fold type of the same.

Topic 2

Is it scientifically meaningful to use structure alignments (via Dali, FATCAT, CE, etc) for protein functional annotation? Discuss the application of this approach and its limitations. Be specific.

As the 3D protein space is extremely challenging to predict and analyze due to the massive amounts of processing power needed to calculate all of the various 3D conformations of proteins, structural alignments for functional annotation are a great way to approximate a protein's 3D structure and extrapolate some hypothesis on functionality from said approximation. Of course, x-ray crystallography is going to be much better at getting the actual 3D structure of proteins, protein bioinformatic tools such as DALI, FATCAT, CE, etc. are all useful in getting a "close enough" answer for many practical needs. One way that I think of the usefulness of protein bioinformatics and other such heuristic tools is that they are useful in getting you in the right neighborhood, but maybe not the exactly correct house. However, it's easier to find the right house if you're somewhat nearby than if you're in a completely different city. While their usefulness in this era of protein bioinformatics cannot be understated, it is highly important that we recognize the shortcomings of using such

tools for protein analysis and annotation. Everything that these tools output are "best guesses" and should be treated as such. For important uses of certain proteins, further investigation and validation are required than just using these tools to find out their approximate structure and functionality, but lab experiments are necessary to go alongside the bioinformatic work to validate any results.

Due date is 18 Nov